Fundamentos de la lingüística de corpus

Concepción de los corpus y métodos de investigación con corpus

Petra Procházková
petra.prochazkova@gmail.com

Texto español revisado por
Tomás Ramírez Minkert
trminkert@yahoo.de

3 de noviembre de 2006

Índice

1. Introducción

La lingüística de corpus trata de la concepción, tratamiento preliminar y análisis de los corpus, y plantea, por ejemplo, qué preguntas lingüísticas se pueden responder por medio del uso de un amplio número de textos. La lingüística de corpus no cumple una función es sí misma, sino más bien se la considera otro método con el cual los lingüistas demuestran sus teorías.

¿Cómo comprueban los lingüistas sus teorías e hipótesis? Para ese propósito, los investigadores tienen cuatro posibilidades: la introspección, los experimentos psicolingüísticos, las encuestas de datos y los corpus. Por medio de la introspección, los lingüistas que tienen la competencia de un hablante nativo confían en su juicio gramatical. Mediante los experimentos psicolingüísticos, los investigadores tienen a disposición, por ejemplo, la medida del tiempo de reacción ante lexemas seleccionados. Por su parte, el procedimiento de encuestas de datos consiste en la elaboración de un cuestionario y de la encuesta de un grupo seleccionado. La tercera posibilidad es recoger datos hasta que se produzca un corpus. De esta manera, el cuarto método se cruza con el tercero.

Se entiende hoy en día como corpus el conjunto de textos que están guardados de forma electrónica y que contienen amplia información lingüística. La unidad más pequeña de un corpus es un „token“ (véase en 4.3).

Alemán	Español
Baumbank/Treebank	base de datos de árboles
Textkorpora	corpus textuales corpus de textos
Repräsentativität	representatividad
Phonetische Merkmale	rasgos fonéticos
Sprachkorpora	corpus orales
Tokenizer	tokenizador
Tokenisierung	tokenización
Parser	parser
Alignierer	alineador
Lemmatisierer	lematizador
Annotationstools	herramientas de anotación
Historische Linguistik	lingüística histórica
Korporaverwendung	aplicaciones de corpus
Monolinguales Korpus	corpus monolingües
Multilinguales Korpus	corpus multilingües
Korpuszusammensetzung	composición del corpus
Monitorcorpus	corpus monitor
Dialektcorpus	corpus dialectal
Multimodale Korpora	corpus multimodales
Referenzcorpus	corpus de referencia
Synchrones Korpus	corpus sincrónico
Diachrones Korpus	corpus diacrónico
herunterladen/downloaden	bajar
Lexikographie	lexicografía
translation memory	memoria de traducción
Historische Korpora	corpus históricos
Tagger	tagger/etiquetador
Korpus	el corpus
Korpora	los corpus
Syntax	sintaxis
Korpora	los corpus/los corpora
Semantik	semántica
Phonologie	fonología
automatische Übersetzung	traducción automática
Computerlinguistik	lingüística computacional
Grenzen von Korpora	límites del corpus
Alignierung	alineación

...continuación de la columna anterior

Alemán	Español
Konkordanzprogramm	programa de concordancias
Parallelkorpus	corpus paralelo
statistisch	probabilístico
Hidden Markov Model	modelos ocultos de Markov
statistische Analyse	análisis estadístico
Bigramme	bigramas
Desambiguierung	desambiguación
Optische Zeichenerkennung	reconocimiento óptico de caracteres
Stichprobe	muestra o prueba al azar o prueba aleatoria
normalverteilt	distribuido normal
Normalverteilung	distribución de Gauss
Grundgesamtheit	universo
Schließende Statistik	estadística inferencial
Fragestellung	enfoque
Wortart	parte de la oración
Lesart	interpretación lectura
Sprechakt	acto de habla

Cuadro 1: Terminología alemán-español

2. Los corpus

2.1. Introducción

Historia

Las colecciones de textos se empleaban ya en el siglo XIX y con anterioridad para describir los cambios en una lengua, para justificar enunciados gramaticales, documentar la adquisición de una lengua, elaborar diccionarios o para hacer comparaciones entre diversas lenguas.
Principalmente se reunían textos de la lengua literaria. Debido a que las evaluaciones tenían que realizarse de forma manual, era difícil efectuar un análisis cuantitativo. Además no había interés en la composición del corpus y en consecuencia los corpus no eran representativos (véase 2.1).
Los primeros corpus electrónicos aparecieron en los años 1960 en forma de tarjetas perforadas y los textos desde luego tenían que ser pasados al ordenador. Hoy en día se puede utilizar el escáner para digitalizar los textos (véase 4.1).

Tipos de corpus

A continuación se describen diversos tipos de corpus:

Bases de datos de árboles Las bases de datos de árboles son textos etiquetados sintácticamente. En general, los análisis sintácticos tienen una estructura en forma de árbol, lo que explica su nombre. Sin embargo, existen también bases de datos de árboles que tienen una estructura de gráfica con conexiones adicionales entre las palabras, en los que la construcción sintáctica no corresponde a un árbol simple, por ejemplo: *NEGRA/TIGER, PDT: Prague Dependency Treebank, Corpus Le Monde, TUT: Turin University Treebank, Spanish Treebank (UAM), ISST: Italian Syntactic-Semantic Treebank, Penn Treebank, Susanne Corpus*

Corpus orales Los corpus orales están constituidos por señales de voz, eventualmente con sus transcripciones de anotación fonética. Un corpus oral contiene grabaciones de llamadas telefónicas, entrevistas o programas de radio, por ejemplo.

Corpus multimodales Los corpus multimodales están constituidos por otros datos orales como prosodia, gestos, movimientos de la boca, inclusive grabaciones sonoros y fílmicas (noticias, documentales).

Corpus de textos Los corpus textuales están constituidos por lengua escrita o por lengua oral transcrita.

Predominan, por lo general corpus textuales que se originan en su totalidad de textos ya que se pueden elaborar con bastante menos esfuerzo que otros corpus. Comúmente, tienen varios cientos de millones de palabras. Otros tipos de corpus cuentan apenas con poco más de un millón de palabras.

Se podría efectuar también otra subdivisión de los corpus:

Corpus sincrónicos vs. diacrónicos Para el corpus sincrónico se recopila material que se compone de la lengua actual (por ejemplo de 1945 hasta el presente). Para el corpus diacrónico se recogen textos de varias etapas históricas de la lengua a fin de poder observar los cambios en la lengua (por ejemplo, los años 1200 hasta 1900).

Corpus monolingües vs. corpus multilingües Los corpus monolingües contienen, en comparación con los corpus multilingües, textos únicamente en una lengua. Los corpus multilingües (ya los corpus bilingües) son muy escasos en comparación con los corpus monolingües porque los textos tienen que existir en versiones traducidas. Los corpus monolingües son de gran utilidad para hacer comparaciones entre las lenguas, elaborar diccionarios y elaborar las memorias de traducción.

Corpus históricos Mientras los corpus textuales modernos pueden recurrir al material ya en forma digital,

los textos para corpus históricos tienen que ser digitalizados por OCR (reconocimiento óptico de caracteres) a través de un escáner. Por ello, deben tomarse en cuenta algunos problemas especiales: ¿Se emplea el manuscrito o una edición? ¿Cómo se manejan las correcciones, las pasadas[1], las glosas, etc.? Otro problema es la codificación de las letras y otros signos de escritura porque algunos caracteres no existen ni siquiera en el Unicode[2].

Corpus de referencia vs. corpus monitor Un corpus de referencia tiene en comparación con un corpus monitor un tamaño establecido, generalmente es de libre acceso y está estandarizado. Por lo contrario, un corpus monitor aumenta de manera constante su tamaño. El corpus incluye, por ejemplo, cada día datos nuevos según criterios fijos, tal como los cumple *Birmingham Bank of English*[3]. El corpus *Wortschatz der Universität Leipzig*[4] presenta cada día las palabras nuevas más frecuentes.

Corpus dialectales Los corpus dialectales están disponibles normalmente sólo en forma oral. Una razón es que los dialectos por lo general no tienen una norma de escritura correcta y que en muchos dialectos no existe en absoluto la tradición de una escritura.

Composición del corpus

Si se quiere establecer un corpus, se tiene que decidir cómo se compone para que sea lo más representativo en atención al enfoque lingüístico. Para eso, el corpus debería cubrir los siguientes parámetros:

lengua oral vs. escrita

diferentes registros especiales: finanzas, medicina, filosofía, gastronomía, electrotecnia

diversos parámetros demográficos: edad, grupo social, género, religión

clasificación por época: (1960-1974, 1975-1993) o (siglo XII-XIV, siglo XV-XVII, siglo XVIII-XX)

diversos medios de comunicación: libros, periódicos, correos electrónicos, radio

diversos niveles lingüísticos: coloquial, formal, familiar, lengua infantil, lengua publicitaria.

diversos tipos de textos: novelas, poemas, formularios, etc.

Representatividad Una desventaja de la mayor parte de los corpus es que no son balanceados en su contenido y por consiguiente los resultados obtenidos no son representativos. Se debe conocer en detalle la composición del corpus con el que se trabaja para saber cuáles cuestiones se pueden responder con el mismo. La razón es que un corpus es siempre sólo un fragmento de una lengua, es decir una prueba aleatoria. Para emitir una declaración válida sobre la "lengua" como un todo al examinar una muestra, se tiene que recurrir a los métodos de la estadística inferencial. La validez de la mayor parte de estos métodos depende en gran medida de si los datos en el universo observado tienen una distribución normal (Gauss). Generalmente, ése no es el caso con los datos de la lengua (por ejemplo, la frecuencia de las palabras). Por ausencia de una muestra para su comprobación como es el caso en una base de datos de árboles no se puede deducir que una cierta construcción gramatical sea incorrecta. Se puede encontrar sólo una prueba PARA una tesis pero nunca por ausencia de las pruebas deducir que alguna tesis sea incorrecta.
¿A partir de qué tamaño es un corpus representativo? A fin de que sea representativo para la experiencia lingüística de una persona, el corpus tendría que contener todos los registros, etc. en las partes porcentuales exactas que el hablante ha recibido en su vida. Un corpus semejante no existe. Aun cuando existiera, sería representativo para la experiencia lingüística de UNA persona en su preciso entorno social, dialecto, año de nacimiento, etc.

[1] Las pasadas son daños en el papel („Rasuren" en alemán).
[2] http://www.unicode.org/
[3] El nombre de COBUILD corpus:Collins Birmingham University International Language Database
[4] http://wortschatz.uni-leipzig.de/

2.2. Aplicaciones de los corpus en la lingüística

Los corpus pueden servir para plantearse numerosos enfoques lingüísticos. En este capítulo se describe cuáles fenómenos lingüísticos buscan las diferentes áreas lingüísticas en un corpus.

Dialectología/Sociolingüística

La dialectología/sociolingüística investiga en el corpus cómo se distinguen los dialectos/sociolectos entre sí y cómo de la lengua estándar (en todos los niveles lingüísticos). Estudia, por ejemplo, cómo cambian los dialectos o si las mujeres hablan de otra forma que los hombres.

Lingüística histórica

La lingüística histórica investiga en el corpus, cuándo una palabra determinada fue reemplazada por otra, cuándo una palabra cambió su modo de uso, cuándo aparece la primera vez una construcción sintáctica, etc.

Psicolingüística

Los psicolingüistas investigan si, por ejemplo, la frecuencia de las palabras influye en la velocidad de reacción y qué papel desempeña la frecuencia de distintas lecturas de las palabras ambiguas. Si el corpus está etiquetado sintácticamente, se puede estudiar qué tamaño tienen las unidades sintácticas del procesamiento y si coinciden con las unidades sintácticas.

Lexicografía

Para crear un diccionario, el lexicógrafo analiza en el corpus en qué contexto aparece una palabra determinada, qué lecturas de una palabra existen, qué palabras aparecen con frecuencia en una combinación estableciada (colocaciones), qué palabras no tienen uso, qué palabras se emplean en un lenguaje profesional o qué sustantivos tienen al mismo tiempo más de un género (fem, mas, neu).

Sintaxis

En la sintaxis se puede verificar con la ayuda de un corpus (sintácticamente etiquetado) si existe una construcción sintáctica. Se puede investigar qué adjetivos aparecen con el verbo *estar* y cuáles con el verbo *ser*, cuáles adjetivos aparecen delante y cuáles detrás de un substantivo o en qué contexto un verbo está en modo subjuntivo y en cuál está en modo indicativo.

Semántica

La semántica léxica estudia en un corpus (etiquetado de lecturas de palabras) cómo se utiliza una palabra determinada y qué sentido tiene (p.ej. *gato*[5]). Además en qué contexto se presentan las lecturas de una palabra o si la palabra aparece como metáfora.

Fonología

La fonología estudia cómo se pronuncian los extranjerismos, si mediante la prosodia es posible distinguir las lecturas de una palabra, cómo se pueden clasificar los acentos y por qué las personas que no son originalmente hispanohablantes cuando hablan español tienen un acento.

Lingüística computacional

Los lingüistas computacionales utilizan los corpus como recurso para la elaboración automática de un diccionario. En la práctica se crearon con métodos cuantitativos el diccionario inglés *Cobuild Dictionary* del corpus *Bank of English* y el diccionario español *Gran Diccionario de Uso del Español Actual*[6] del corpus *Cumbre*[7]. Además, los lingüistas computacionales usan los corpus como recurso para la extracción de frases de los corpus bilingües para una memoria de traducción, para la extracción automática de colocaciones, para la extracción automática de las diferencias

[5] El gato: *Katze/Wagenheber* en alemán.
[6] http://www.um.es/lacell/proyectos/diccionario/
[7] http://liceu.uab.es/~joaquim/language_resources/lang_res/-Corp_text_esp.html

del lenguaje en todos los niveles (sintaxis, semántica, etc.) entre, por ejemplo, el español peninsular y el español en los Estados Unidos.

Otras Aplicaciones

En un corpus se pueden buscar los errores típicos o se puede investigar cómo los niños aprenden las normas ortográficas. Con la ayuda de la estilometría se puede reconocer de qué autor es un texto determinado o se puede examinar el procesamiento de adquisición de la lengua por parte de los niños en distintas edades. En un corpus compuesto de textos escritos en español por personas no hispanohablantes, se pueden analizar los errores característicos que cometen. En la morfología se puede estudiar qué sufijos derivados existen y cuáles de ellos son productivos. La pragmática examina, por ejemplo, cuáles posibilidades existen en español para señalizar al interlocutor que queremos poner a fin la conversación.

2.3. Búsqueda en los corpus

Herramientas de búsqueda

Las herramientas de búsqueda denominadas programas de concordancia suelen mostrar los resultados en el formato *kwick* (keyword in context - palabra clave en contexto). Es decir, que muestran la palabra buscada en el contexto utilizado y se puede determinar también la longitud del contexto a ambos lados de la palabra clave. Por ejemplo, se desarrollaron los programas *Sara*[8] para buscar en el corpus inglés BNC, *Cosmas*[9] para buscar en el corpus alemán en IDS y *Bonito*[10] para buscar en el corpus checo *CNK*[11]. Para investigar en los corpus españoles está a disposición un programa web[12] en el Departamento de Lenguas Romances en la Humboldt Universidad de Berlín[13]. Programas complejos como *TigerSearch*[14] también permiten visualizar la estructura de la frase encontrada.

Proceso de búsqueda

La posibilidad más fácil que ofrecen todas estas herramientas es la búsqueda exacta por un token. Por ejemplo se puede buscar por los tokens siguientes: *hispano*, *hispana*, *hispanos*, *hispanas*. Si el corpus está lematizado, se puede buscar también por el lema *hispano* para encontrar las formas antes nombradas.
Una función avanzada de estas herramientas es la búsqueda por tokens con la ayuda de expresiones regulares. Por ejemplo, la expresión $[H \mid h]\, ispan.*$ encuentra todos tokens que comienzan con *H* o con *h* seguidos de la cadena de caracteres *ispan* y luego seguida de cualesquiera caracteres. De esta manera, se localizan también *hispanoparlantes*, *hispanoamericanos* o *hispano-argentina*.
Si el corpus está etiquetado con partes de la oración (en inglés *pos*: part of speech) y si la herramienta permite buscar también por la secuencia de las palabras, se pueden encontrar, por ejemplo, todos los adjetivos que siguen al lema *estar*.

Con muy poca frecuencia es posible buscar semánticamente por lecturas de una palabra, como el caso de la palabra *ganar* en los sentidos diferentes: *en el juego, mejorar, trabajar, llegar a, aventajar* o buscar por rasgos fonéticos, por ejemplo, cómo una persona pronuncia *paella*. En muy pocas ocasiones es posible también buscar por informaciones pragmáticas (p.ej. actos de habla) por falta de anotación pragmática de corpus.

Límites del corpus/de la búsqueda

Con la ayuda de un corpus no se puede determinar la etimología de las palabras, no se puede establecer si una construcción es incorrecta, no se puede descubrir qué extensión tiene la frase más larga que se puede entender y por medio de un corpus no se puede determinar qué sentido tiene una frase. Por lo general, se puede buscar solamente lo que está etiquetado, por ejemplo, las informaciones semánticas de las pa-

[8] http://www.natcorp.ox.ac.uk/tools/
[9] http://www.ids-mannheim.de/cosmas2/
[10] http://ucnk.ff.cuni.cz/bonito/
[11] http://ucnk.ff.cuni.cz/
[12] http://rom99.sprachen.hu-berlin.de/latinus/korpora/login.php
[13] http://www2.hu-berlin.de/romanistik/
[14] http://www.ims.uni-stuttgart.de/projekte/TIGER/TIGERSearch/

labras o las meta-informaciones como el autor de un texto, la fecha del texto, y otros.

3. Corpus disponibles

3.1. Los corpus más conocidos

Los corpus más conocidos del mundo son *Brown corpus*[15] (en), *British National Corpus*[16] (en), *Bank of English*[17] (en), *London-Lund Corpus* (en), Lancaster-Oslo-Bergen Korpus[18] (en), *Tycho Brahe Parsed Corpus of Historical Portuguese*[19] (pt), *Český národní korpus*[20] (cz), *Childes*[21], *MULTEXT*[22], *ECI*[23] (multilingüe), *Projekt Gutenberg*[24] (de, en) y los IDS corpus[25](de).

3.2. Corpus españoles

CREA

El *Corpus de Referencia del Español Actual*[26] contiene más de 200 millones de tokens, 10 % de los datos son transcripciones de la lengua oral y 50 % de los datos provienen de España. El corpus cubre el período de 1975 hasta ahora en la lengua. Fue desarrollado por el *Instituto de Lexicografía de la Real Academia de la Lengua Española*[27] y se compone de los subcorpus siguientes:

Corpus Oral de referencia del español contemporáneo

ACUAH Análisis de la Conversación de la Universidad de Alcalá - de Henares

ALFAL Macrocorpus de la norma lingüística culta de las principales ciudades del mundo hispánico, de la Asociación de Lingüistica y Filología de América Latina.

Caracas-77 Estudio Sociolingüístico de Caracas, 1977

Caracas-87 Estudio Sociolingüístico de Caracas, 1987

CEAP Corpus de Encuestas de Asunción de Paraguay

COVJA Corpus oral de la variedad juvenil universitaria del español hablado en Alicante

CSC Corpus para el estudio del español hablado en Santiago de Compostela

CSMV Corpus Sociolingüístico de Mérida - Venezuela, 40 horas de entrevistas con 80 hablantes maternos, Universidad de los Andes.

UAM Corpus Oral de Referencia del Español Contemporáneo

PÚBLICO Material público procedente de Internet.

CORDE

El *Corpus Diacrónico del Español*[28] de RAE. Es una colección de las primeras fuentes españolas hasta ahora, contiene 70 millones de tokens.

CUMBRE

El *Corpus lingüístico del español contemporáneo*[29] contiene 70 milliones de tokens. El 50 por ciento de los datos proviene de España. Consiste de corpus textuales que reúnen datos desde los años de 1950 y de los corpus orales de dos millones de tokens de los años 1990.

[15] Brown Corpus of Standard American English: http://en.wikipedia.org/wiki/Brown_Corpus
[16] BNC: http://www.natcorp.ox.ac.uk/
[17] http://www2.lingsoft.fi/doc/engcg/Bank-of-English.html
[18] http://khnt.hit.uib.no/icame/manuals/LONDLUND/INDEX.HTM
[19] http://www.ime.usp.br/ tycho/corpus/files/index.html
[20] http://ucnk.ff.cuni.cz/
[21] http://childes.psy.cmu.edu/
[22] http://aune.lpl.univ-aix.fr/projects/multext/, http://nl.ijs.si/ME/
[23] http://www.elsnet.org/eci.html
[24] http://gutenberg.spiegel.de/
[25] http://corpora.ids-mannheim.de/
[26] http://corpus.rae.es/creanet.html, http://www.lllf.uam.es/corpus/corpus.html
[27] http://www.rae.es/
[28] http://corpus.rae.es/cordenet.html
[29] http://www.lllf.uam.es/~fmarcos/informes/corpus/corpulee.html, ftp://ftp.lllf.uam.es/pub/corpus/oral/

ARTHUS

El *Archivo de textos hispánicos de la Universidad de Santiago*[30] contiene corpus textuales y orales de diferentes épocas de la historia española en España e Hispanoamérica. Está etiquetado sintácticamente.

Base de Datos Sintácticos del Español Actual

Contiene 160.000 cláusulas que componen la parte contemporánea del ARTHUS [31]

UAM-Treebank

La base de datos de árboles *UAM Spanish Treebank*[32] es un corpus de textos escritos compuesto por 1600 oraciones extraídas de los periódicos digitales El País[33] y Compra Maestra.

Chile

El *Corpus lingüístico de referencia de la lengua española en Chile*[34] consiste de dos millones de tokens.

Argentina

El *Corpus lingüístico de referencia de la lengua española en Argentina*[35] consiste también de dos millones de tokens.

MC-NLCH

El *Macrocorpus de la norma lingüística culta de las principales ciudades del mundo hispánico* se compone de la transliteración de ochenta y cuatro horas de grabación (de 12 ciudades del área hispanohablante).

[30] http://gramatica.usc.es/EspArthus.html
[31] http://www.bds.usc.es/
[32] http://www.lllf.uam.es/ sandoval/UAMTreebank.html
[33] http://www.elpais.es/
[34] http://www.lllf.uam.es/~fmarcos/informes/corpus/cochile.html, ftp://ftp.lllf.uam.es/pub/corpus/chile/
[35] http://www.lllf.uam.es/~fmarcos/informes/corpus/coarginl.html, ftp://ftp.lllf.uam.es/pub/corpus/argentina/

Corpus españoles como parte de un corpus multilingüe

El corpus *Multext*[36]: un millón de tokens.

El corpus *ECI*[37] con los subcorpus: el diario *Sur*[38], *XeroX ScanWorx User's Guide*, *Reports of the Committee on Freedom of Association of the Governing Body of the ILO and related material*, *The announcement text of the EC Esprit program*.

El corpus *Childes*[39].

Las leyes y protocolos de la Unión Europea.

El corpus *CRATER*[40].

El Corpus *ITU*[41] un millón de tokens.

El Corpus *OPUS*[42]: documentación de OpenOffice y php

La *Declaración Universal de Derechos Humanos*[43]

El Corpus *JRC-ACQUIS*[44]: textos jurídicos de la UE.

El *diario oficial de la UE*[45]

Corpus para analizar las variedades geográficas

ALMECOR Universidad de Granada. Cuatro zonas, 30 horas de grabación.

FAE_Esp Can Fonética acústica y experimental del español de Canarias. Universidad de Laguna, 30 minutos por hablante.

ILSE[46] *El Corpus del Habla en Almería*, 75 horas.

VUA Variedades urbanas andaluzas. Universidad de Granada y de Málaga, 250 horas con 290 hablantes.

[36] http://aune.lpl.univ-aix.fr/projects/multext/
[37] ECI: European Corpus Initiative
[38] http://www.diariosur.es/
[39] http://childes.psy.cmu.edu/
[40] CRATER: Corpus paralelo sobre telecomunicaciones, http://www.comp.lancs.ac.uk/linguistics/crater/corpus.html
[41] ITU: International Telecommunications Union
[42] http://logos.uio.no/opus
[43] http://www.unhchr.ch/udhr/index.htm
[44] http://wt.jrc.it/lt/Acquis/
[45] http://europa.eu.int
[46] http://www.grupoilse.org/corpus/corpus22.htm

Corpus para el análisis del discurso

ADPA Universidad de La Coruña, 75 horas.

Corpus para el análisis de la evolución de la lengua

Adquisición, desarrollo y representación de categorías semánticas en niños de edad escolar de UNED. Cerca de veinte mil tokens, 846 frases.

Diferencias individuales en la adquisición del lenguaje Univ. de Barcelona, diez hablantes, diez sesiones al año, 3-4 años.

Corpus de habla infantil de CSIC-UNED. seis hablantes con 15-20 sesiones al año. Cada año desde 1991. Cada sesión dura 45 minutos.

Disponibilidad léxica de los adolescentes Univ. de Salamanca, 200 mil tokens.

Otros corpus

Corpus del español[47] (Illinois State University)

Periódicos: *ABC*[48] en el CD con más de cuatro millones de tokens, *El Mundo*[49] 1994-1995 en el CD.

Briscoe: Corpus etiquetado de 17.000 tokens.

El Corpus Virtual de la red[50].

Corpus de transcripción de la lengua oral para el estudio de la norma lingüística culta de la lengua española hablada en Madrid.

Corpus 92 Lengua Escrita por aspirantes a estudios universitarios. Universidad Pompeu Fabra, dos millones de tokens.

Corpus Textual del Español Periodístico.

LAN textos técnicos de la empresa Micro Focus SA. 26 mil tokens.

Archivo Digital de Manuscriptos y Textos Españoles en el CD. Dept. of Spanish & Portuguese, U. of California, Berkeley.

Corpus de vocabulario del niño de 6 a 14 años Diccionario de frecuencias. Universidad de Granada.

Corpus contrastivo español/francés Universidad de Sevilla. Análisis comparativo posible de errores de traducción.

LEGEBIDUN[51], español-vasco, Universidad de Deusto.

LEJES Proyecto de las Universidades de Bonn y de Granada. Análisis de palabras jurídicas, cinco mil tokens.

Camtie y otros en VISL[52].

4. Generación de los corpus

4.1. Adquisición de datos para un corpus

World Wide Web La red ofrece una cantidad inmensa de páginas web que se puedan bajar automáticamente, por ejemplo, con el programa BootCat[53]. Principalmente para componer un corpus de una lengua minoritaria se usa ese método si no hay ninguna o sólo pocas fuentes de esta lengua.

Escáner Se escanean libros y otro textos y se aplica el reconocimiento óptico de caracteres para obtener un texto en forma electrónica. El reconocimiento óptico no es perfecto, hay que corregir el resultado. Además hay que borrar los números de páginas y bordes de textos escaneados y hay que reconocer dos columnas en una página como en este trabajo y no tratar dos columnas como una sola.

[47] http://www.corpusdelespanol.org
[48] http://www.abc.es/
[49] http://www.elmundo.es/
[50] www.gelbukh.com/CV/Publications/2002/MLIA-2002-Corpus.pdf
[51] http://www.serv-inf.deusto.es/abaitua/konzeptu/lege2dun.htm
[52] http://corp.hum.sdu.dk/cqp.es.html
[53] http://sslmit.unibo.it/~baroni/bootcat.html

Grabación de tono Para obtener datos orales se hacen grabaciones de los medios de comunicación (radio, películas, talkshows), se graban conversaciones de la vida cotidiana (calle, escuela) o se graban entrevistas y conversaciones de las operaciones quirúrgicas o conversaciones en la cabina del piloto.

Compra La posibilidad más cómoda y corriente de obtener datos para un corpus es la compra del archivo en la versión electrónica de un diario o de un periódico, por ejemplo, *El País*[54]. Si el corpus obtenido no está en formato binario se puede tokenizar, lematizar e etiquetar con más informaciones lingüísticas - con el resultado de que se puede buscar no sólo por una palabra.

4.2. Anotación de un corpus

Con la ayuda de elementos llamados **tag**s (*<algún tag>*) se puede enriquecer el texto con informaciones estructurales. Las herramientas de anotación disponibles determinan cuánto se requiere para efectuar este proceso.

Principios de la anotación

Las anotaciones se deben codificar de manera tal que se pueda rehacer el texto original. La regla establece que el contenido está separado de la estructura/información meta (difícil con la anotación fonética).

La evaluación de las anotaciones debe ser posible sin el texto original.

Las normas de anotación deben ser accesibles.

Los anotadores y las circunstancias de la anotación deben ser conocidos.

Los usuarios deben saber que las anotaciones pueden contener errores.

Los investigadores deben mantener una anotación que sea neutral ante las teorías (difícilmente posible)

Se debe tomar en cuenta estándares de codificación TEI[55] y CES[56] de ELRA[57], LDC[58] y EAGLES[59].

El formato de la anotación

Para lograr el propósito de separar el contenido de la estructura se utilizan lenguajes de marcación (markup languages) como HTML[60], SGML[61] o XML[62].

HTML[63] es inadecuado para la anotación porque el conjunto de los tags es limitado (<p>,
,<h1> y algunos más) y además no cumple la norma que establece que a un tag abierto sigue también un tag cerrado.

SGML[64] ofrece la posibilidad de anotación más amplia, aunque al mismo tiempo es la opción más costosa y por ello, solamente útil para proyectos grandes (por ejemplo, el corpus CREA).

XML[65] es la lengua más adecuada para la codificación porque comparado con HTML puede contener un conjunto de tags infinito. Al mismo tiempo es mucho más fácil en su estructura como SGML. La organización CEI[66] propuesta con una DTD[67] especial sobre cómo se debía codificar un texto.

Herramientas de anotación

MMAX[68]: Para la anotación multimodal.

[54] http://www.elpais.es/

[55] TEI: Text Encoding Initiative.
[56] CES: Corpus Encoding Standard Encoding Initiative.
[57] ELRA: European Languages Ressources Agency.
[58] LDC: Linguistic Data Consortium.
[59] EAGLES: Expert Advisory Group on Language Engineering Standards.
[60] HTML: Hypertext markup language.
[61] SGML: Structured generalized markup language.
[62] XML: Extensible markup language, http://www.w3.org/XML/
[63] HTML: Hypertext Markup Language.
[64] SGML: Standard Generalized Markup Language.
[65] XML: Extensible Markup Language
[66] CEI: Corpus Encoding Initiative.
[67] DTD: Document Type Definition.
[68] http://www.eml-research.de/english/research/nlp/download/mmax.php

NITE XML[69]: Para la anotación multimodal.

@*nnotate*[70]: Para la anotación sintáctica

EXMARaLDA[71]: Para la anotación de discurso.

PALinkA[72]: Para la anotación de discurso.

Transcriber[73]: Para la anotación fonética.

Praat[74]: Para la anotación fonética.

Anvil[75]: Para la anotación de videos.

Elan[76]: Para la anotación de videos.

TASX[77]: Para la anotación de videos.

4.3. Procesamiento de un corpus

Para elaborar un corpus hay que transformar con anterioridad el texto original en una forma en la que posteriormente se pueda acceder al corpus y extraer del mismo la mayor cantidad de informaciones posibles. Por eso el texto se tokeniza, etiqueta, parsea y alínea.

Tokenizador

Tokenizador es un programa que tiene la tarea de segmentar un texto en tokens. Por token se entiende una cadena de caracteres y cifras que están encerradas entre espacios, en la situación ideal. El tokenizador incluye también una segmentación más detallada, por ejemplo, reconoce y divide la cadena de caracteres *"soler"* en tres tokens para poder etiquetar los tres y permitir que se pueda buscar la palabra *soler*. El tokenizador reconoce los signos de puntuación y los separa de su entorno. Incluso, tokenizador eficiente reconoce como un todo los números del teléfono *(012) 34 567 89*. El tokenizador tiene problemas para reconocer *Euro 40,-* como un todo o reconocer siglas que terminan con un punto y se encuentran al final de una frase.

Tagger/etiquetador

Un tagger es un programa que asigna a cada token automáticamente un *tag*: su parte de oración (pos) y su lema (a veces también otras informaciones morfosintácticas). La cantidad de diferentes *pos* que asigne un tagger depende del *tagset* (conjunto de tags) del programa - cambia entre 10 y 1000 diferentes pos-tags. Un tagset español - propuesto por EAGLES - distinque 250 pos-tags. Existen etiquetadores probabilísticos y aquellos que están basados en reglas.

Asignación de un tag El tagger revisa en su léxico electrónico y asigna al token el correspondiente tag. Si el token es ambiguo, como por ejemplo, *sobre* (puede ser una preposición o un sustantivo) el tagger decide por el contexto cuál tag debe asignar. Problemas con la asignación: Palabras que no están en el léxico (extranjerismos, palabras creadas espontáneamente), palabras de dos componentes como en *Los Angeles* o palabras con errores ortográficos. Ambigüedades típicas en un corpus español son, por ejemplo, *creo* (*creer, crear*), *para* (*parir, parar, para*), *casa* (*casar, casa*), *nada* (*nadar, nada*), *parte* (*partir, parte*), *sentido* (*sentir, sentido*), *fui* (*ser, ir*), *ayuda* (*ayudar, ayuda*).

Parseador

Los parseadores se utilizan en la lingüística de corpus para etiquetar sintácticamente las frases para la base de datos de árboles. Debido a que no existen parseadores que analizan correctamente una frase compleja, los parseadores se limitan a reconocer *chunks* (sintagmas) que después se corrigen de forma manual y se los coloca juntos para componer de nuevo la frase. Este trabajo lo efectúan dos personas, por eso, la calidad de las bases de datos de árboles es alta, por el contrario el tamaño del corpus es pequeño.

[69] http://www.ltg.ed.ac.uk/NITE/
[70] http://www.coli.uni-saarland.de/projects/sfb378/negra-corpus/annotate.html
[71] http://www1.uni-hamburg.de/exmaralda/
[72] http://clg.wlv.ac.uk/projects/PALinkA/
[73] http://trans.sourceforge.net/en/presentation.php
[74] http://www.fon.hum.uva.nl/praat/
[75] http://www.dfki.de/ kipp/anvil/
[76] http://www.mpi.nl/tools/elan.html
[77] http://medien.informatik.fh-fulda.de/tasxforce

Alineador

El alineador es un programa que se encarga de la asignación de elementos traducidos (palabras, sintagmas, frases). Una alineación es necesaria para obtener corpus paralelos.

Taggers españoles

dtt[78] (decision tree tagger)

wraetlic-tagger[79] (wraetlic-tools)

svm-tagger[80]: support vector machine tools

etiproct-tagger: Etiquetador y procesador de Corpus.

freeling-tagger[81]: FreeLing

VISL-Tagger[82] de Syddansk Universitet, online.

5. Conclusión

Desde el surgimiento de Internet, numerosos textos electrónicos están disponibles en la red para elaborar un corpus. Al mismo tiempo, los ordenadores han sido lo suficientemente versátiles para procesar y archivar los textos. Además, las herramientas de anotación aceleran hoy en día el proceso para la creación de un corpus. De este modo, cualquier científico puede elaborar un corpus grande en corto tiempo para investigaciones lingüísticas. Sólo hay que tener en cuenta los derechos de autor en la composición y la transferencia del corpus.

Agradecimiento

Deseo expresar mi sincero agradecimiento a Tomás Ramirez Minkert por su grandísima ayuda en las correcciones ortográficas y en el estilo del presente artículo. Sin la revisión del texto, este trabajo no sería posible.

[78] http://www.ims.uni-stuttgart.de/projekte/corplex/TreeTagger/
[79] http://www.ii.uam.es/ ealfon/eng/research/wraetlic.html
[80] http://www.cs.wisc.edu/dmi/svm/
[81] http://garraf.epsevg.upc.es/freeling/
[82] http://visl.sdu.dk/

6. Referencias

[1] Biber, Douglas; Conrad, Susan & Reppen, Randi (1998) *Corpus Linguistics. Investigating Language Structure and Use.* Cambridge University Press, Cambridge.

[2] Evert, Stefan & Fitschen, Arne (2001) *Textkorpora*. In Carstensen et al.(2001) S. 369-376.

[3] Kennedy, Graeme (1998) *An introduction to corpus linguistics.* London [u.a.]: Longman.

[4] McEnery, Tony & Wilson, Andrew (2001) C*orpus Linguistics.* Edinburgh University Press, Edinburgh.

[5] Sinclair, John (1995) *Corpus, concordance, collocation.* Oxford [u.a.]: Oxford Univ. Press.

[6] Baayen, Harald (2001) *Word Frequency Distributions.* Kluwer, Dordrecht.

[7] Feldweg, Helmut und Erhard W. Hinrichs [Eds.] (1996) *Lexikon und Text: wiederverwendbare Methoden und Ressourcen zur linguistischen Erschließung des Deutschen.* Tübingen: Niemeyer.

[8] Garside, Roger; Leech, Geoffrey & McEnery, Tony [Eds.] (1997) *Corpus Annotation: Linguistic Information from Computer Text Corpora.* Addison Wesley Longman, New York.

[9] Carstensen, Kai-Uwe et al. [Eds.] (2004) *Computerlinguistik und Sprachtechnologie. Eine Einführung.* Spektrum Akademischer Verlag, Heidelberg.

[10] Svartvik, Jan [Ed.] (1992) *Directions in Corpus Linguistics.* Mouton de Gruyter, Berlín.

[11] Pusch, Claus D., Wolfgang Raible [Eds.] (2002) R*omanistische Korpuslinguistik. Korpora und gesprochene Sprache.* Gunter Narr Verlag, Tübingen.

[12] Kennedy, Graeme (1998) *An introduction to corpus linguistics.* London: Longman.

[13] Garside, Roger & Leech, Geoffrey (1997) *Corpus Annotation: Linguistic Information from Computer Text Corpora.* New York: Addison Wesley Longman.

[14] Pusch, C. D. & Kabatek, J. & Raible, W. (2005) *Romanistische Korpuslinguistik*. Narr.

[15] Baldry, Anthony & Thibault, Paul (2005) *Multimodal Transcription and Text Analysis.* London: Equinox Publishing Ltd.

[16] McEnery, Anthony & Xiao, Richard & Tono, Yukio (2006) *Corpus-Based Language Studies: An Advanced Resource Book*. Routledge.

[17] Lüdeling, A & Kyto, M. & McEnery, E (Eds.) (en preparación) *Handbooks of Linguistics and Communication Science Volume Corpus Linguistics.* Berlin: Mouton de Gruyter.